

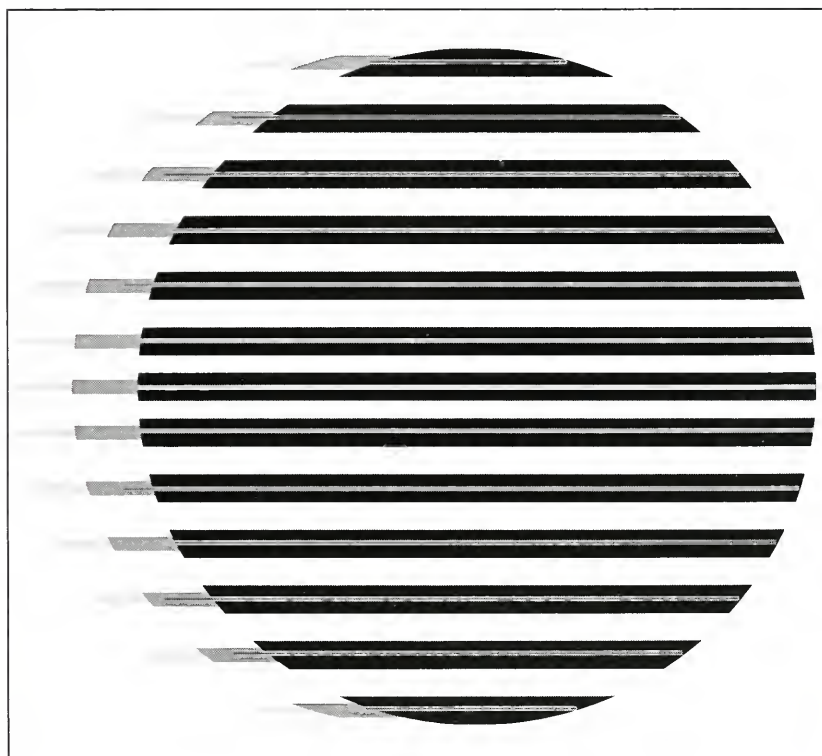
# IASSIST


Q U A R T E R L Y

VOLUME 22

Fall 1998

NUMBER 3





Digitized by the Internet Archive  
in 2010 with funding from  
University of North Carolina at Chapel Hill

<http://www.archive.org/details/assistquarterly223inte>

# IASSIST QUARTERLY

The IASSIST QUARTERLY represents an international cooperative effort on the part of individuals managing, operating, or using machine-readable data archives, data libraries, and data services. The QUARTERLY reports on activities related to the production, acquisition, preservation, processing, distribution, and use of machine-readable data carried out by its members and others in the international social science community. Your contributions and suggestions for topics of interest are welcomed. The views set forth by authors of articles contained in this publication are not necessarily those of IASSIST.

## Information for Authors:

The QUARTERLY is published four times per year. Authors are encouraged to submit papers as word processing files. Hard copy submissions may be required in some instances. Word processing files may be sent via email to [jtstratford@ucdavis.edu](mailto:jtstratford@ucdavis.edu). Manuscripts should be sent to Editor: Juri Stratford, Government Information and Maps Department, Shields Library, University of California, 100 North West Quad, Davis, California 95616-5292. Phone: (530) 752-1624.

The first page should contain the article title, author's name, affiliation, address to which correspondence may be sent, and telephone number. Footnotes and bibliographic citations should be consistent in style, preferably following a standard authority such as the University of Chicago press *Manual of Style* or Kate L. Turabian's *Manual for Writers*. Where appropriate, machine-readable data files should be cited with bibliographic citations consistent in style with Dodd, Sue A. "Bibliographic references for numeric social science data files: suggested guidelines". *Journal of the American Society for Information Science* 30(2):77-82, March 1979. Announcements of conferences, training sessions, or the like, are welcomed and should include a mailing address and a telephone number for the director of the event or for the organization sponsoring the event.

## Editors

### Karsten Boye Rasmussen,

Eckersbergsvej 56,  
5230 Odense M,  
Denmark.  
Phone: +45 6612 9811,  
Email: [kbr@sam.sdu.dk](mailto:kbr@sam.sdu.dk)

### Juri Stratford

Government Information and  
Maps Department,  
Shields Library,  
University of California,  
100 North West Quad,  
Davis, California 95616-5292  
Phone: (530) 752-1624.  
Email: [jtstratford@ucdavis.edu](mailto:jtstratford@ucdavis.edu)

## Production

### Laura Bartolo,

Libraries and Media  
Services,  
Kent State University,  
Ohio 44242.  
Phone: (330) 672-3024, x31.  
Email:  
[lbartolo@kentvm.kent.edu](mailto:lbartolo@kentvm.kent.edu)

### Walter Piovesan

Research Data Library  
Simon Fraser University  
Burnaby, B.C.  
Canada V5A 1S6.  
Phone: (604) 291-5937.  
Email: [walter@sfu.ca](mailto:walter@sfu.ca)

# C O N T E N T S

Volume 22

Number 3

Fall 1998



## FEATURES

- 4 Data liberation, bridges to cross  
*Richard Baily*
- 9 The California Digital Library:  
*Implications for Social Science Data  
Files Collections*  
*Daniel C. Tsang*
- 12 Privacy and Computerized Data Bases  
*Thomas E. Brown*
- 17 Data Protection and Privacy in the United  
States and Europe  
*Jean Slemmons Stratford & Juri Stratford*

Title: Newsletter - International Association for Social  
Science Information Service and Technology

ISSN - United States: 0739-1137 © 1998 by IASSIST. All  
rights reserved.

# Data liberation, bridges to cross

## Abstract

In Canada, the use of statistical data (micro data files and major databases) for teaching and research is an important phenomena that does not seem to be losing strength in the near future. This situation is a major consequence of the Data Liberation Initiative (DLI), established in 1996 as a partnership among Statistics Canada, other federal departments and Canada's academic community. The idea of providing affordable access to Canadian information results from a co-operative effort among the Humanities and Social Science Federation of Canada (HSSFC), the Canadian Association of Research Libraries (CARL), the Canadian Association of Public Data Users (CAPDU) and the Canadian Association of Small University Libraries (CASUL). Less than two years after its inception more than 50 universities have joined the consortium, which is a clear indication of a true willingness to make data more available.

This illustrates the fact that the high cost of buying data was an obstacle to its availability, especially in small universities where the absence of a minimum number of students results in a higher cost/benefit ratio related to data acquisition. However, there are still many obstacles to free numerical data use. If some Canadian universities have a long history in data services (Carleton University's Data Centre celebrated its 30th anniversary in 1996), such a tradition does not exist everywhere, especially in small universities.

To maximise use of data files, increased education at the reference staff level and at the consumer level, including professors, must occur. Data usage requires a good knowledge of data extraction and associated analytical instruments. How can these tools be made accessible to customers who are not able to manipulate data files, but who have a definite need for the information? How can we satisfy different needs for different types of users? How can data be included in the academic curriculum? How can data librarians play their educational role and how can this role be balanced with professorial responsibilities? Fortunately, interesting answers are unfolding.

by Richard Boily \*

## Introduction

Numerical data collected from various surveys conducted throughout the nation by organisations such as Statistics Canada constitute an information source that is both important and extremely powerful in understanding social phenomena. Access to these data is necessary for teaching and academic research. In fact, this access is essential to intellectual freedom and democracy.

In Canada, the conditions of accessibility to numerical data have undergone major changes within the past two past years, due to the implementation of the Data Liberation Initiative (DLI) by Statistics Canada.

The Data Liberation Initiative is a management framework that modifies the conditions of data access. These changes coincide, and certainly not by accident, with the arrival of new technologies: the development of both powerful personal computers and their increased data storage capacity and of user-friendly software (Excel and SPSS), in addition to the advent of the Internet. Such events bear directly on the theme of this conference, notably Global Access and Local Support. With new parameters defined by DLI, Canadian data are potentially more accessible than ever to Canadians.

Even if DLI is successful, the fact remains that widespread numerical data use in Canadian universities is uncommon and many obstacles exist that would allow the situation to change. The objective of this presentation is to examine problems of accessibility to numerical data within the context of DLI.

This presentation is composed of three parts. First, a brief history of the origins of DLI and its role will be given. This will also entail a review of the objectives. Then, the problems of developing numerical data use within the context of DLI will be addressed. Finally, it will be shown that there are elements of DLI that represent an opportunity to improve the democratisation (accessibility) of data in Canada.

## 1.DLI

### Origins

Traditionally, Statistics Canada publishes the statistical information that it has collected in the form of aggregated data tables. These documents are largely distributed to libraries via the government publication deposit program. However, numerical data files that have been excluded from the deposit program and until recently, were available only at a very high price. Such a situation has been strongly discredited by the research community, notably by Professor Paul Bernard, professor of sociology at the University of Montreal and a member of the National Statistics Council. In 1991, Professor Bernard asserted that, "the genuine exercise of democracy increasingly requires that citizens get access to complex information and have the skills required to understand it".

In 1993, following the opinions voiced by Professor Bernard and others, several individuals representing the Social Sciences and Humanities Research Council of Canada, the Association of Universities and Colleges of Canada, the Canadian Association of Research Libraries and the Canadian Association of Public Data Users united under the auspices of the Social Science Federation of Canada. Their specific objective was to develop a strategy to render Canadian survey data more accessible to the research and teaching community. The work of this group led to a proposal that rapidly passed through the various levels of the federal government and thus, was accepted by Statistics Canada. The DLI received official recognition from the Treasury Board of Canada in February 1996. It was subsequently included as part of the Canadian government's Science and Technology Strategy in March of the same year.

### What is the DLI?

The DLI is a five-year project among universities, Statistics Canada and several federal departments. Under the agreement, participating universities pay a known and affordable yearly fee (\$12,000 for CARL members or \$3,000 for CASUL members), that gives them access to all standard data products provided by Statistics Canada. FTP on the Internet is the primary method of accessing these files. If files exist only on CD-ROM each participant is entitled to a copy. However, if they exist in both forms, users may choose one or both types of files. Participating libraries must make acquired data available to their users while, at the same time, insuring that they do not use it for commercial purposes.

### Objectives of the DLI.

As stated on the DLI Web site, itself, "timely access to data

is essential if researchers are to focus on Canadian problems and students are to learn to analyse Canadian information. Without affordable data for research and training, Canada risks producing innumerate graduates and basing its policy decisions on incomplete information. Independent analyses enhance public debate and policy making on questions relevant to all Canadians. The federal government invests large amounts of public money in data collection. The DLI can ensure a valuable return on this investment by distributing data to the university community, which will encourage analysis and put more information in the public domain".

Before the inception of DLI, we experienced the embarrassing situation where Canadian researchers, who needed to develop methodological expertise or study a particular social phenomena, had to work with American data because Statistics Canada data were either too expensive or not available at geographically specific levels. Unfortunately, this situation still exists.

### 2.Numerical data use in the context of DLI, or, how do things happen now?

Although DLI has been in existence barely two years, it is

Table 1. University participation to DLI

Universities offering data services before DLI	Participating universities in the Data Liberation Initiative (DLI)		
	1996	1997	1998
Between 15 and 20	50	59	61

showing positive results that are measurable and it seems to be fulfilling expectations.

There is an excellent participation among Canadian universities in DLI that exceeds even the most optimistic expectations. Before DLI's inception, barely 15 to 20 universities offered any form of numerical data service. This obviously does not take into account individual professors and researchers who ordered files from Statistics Canada. It is even probable that the acquisition of these data has been made through the library. It does not take account either of the various numerical data files, notably the 1986 and 1991 Canadian census data distributed on CD-ROM, that were made available by several libraries well before the inception of DLI. For the past several years, we have offered training sessions on CD-ROM census data search at our library. However, the fact remains that these transactions were not integrated within a real data service.

Today, more than 60 institutions participate in the DLI consortium, which is nearly 80% (61/78) of Canadian

universities. Fifty of those affiliated have been with DLI since its first year. However, it does not follow that all 61 participating institutions offer a numerical data service. This number simply indicates that there is, at minimum, a DLI representative in each of these universities and that this person is involved, to one degree or another, in data-related activities that may lead to the implementation of a data service.

Canadian society to be conducted. This is the real meaning of accessibility and the democratisation of data.

### Context of numerical data use

Even if Statistics Canada survey data are potentially available (data can be downloaded at any given time by whoever needs it) and the price no longer constitutes an obstacle to use, both the intrinsic complexity of the data and the means of exploiting it remain major constraints for its use. The majority of users are, in fact, incapable of manipulating raw data. Their need for a data service before the arrival of massive data sets occurs is more important than ever. However, the costs associated with setting-up and maintaining such a service are considerable and largely exceed

the cost of the data. These costs constitute a major obstacle to the democratisation of data.

### Variety of resource persons

The participation of new institutions in DLI and, the consequent arrival of new representatives are promising events for the development of new data services. On the other hand, all those who have accepted to be responsible for data files do not have, necessarily, the same level of competence. Not all have the same interest or desire to develop the expertise required by this new function.

The DLI is a young program that has experienced accelerated development (50 members the first year), probably due to a copy-cat effect. As we have previously stated, it is necessary to remember that statistical information distributed by Statistics Canada is traditionally published first as working documents and then, these documents are often integrated into governmental publications. For librarians who specialise in governmental publication reference, the responsibility of numerical data management constitutes a sizeable challenge. In many cases, and I have to recognise that it was true with me, new DLI representatives came to the job previously unaware of what was entailed with numerical data files.

Numerical data exploitation and use, especially micro data, supposes a knowledge of computers and statistics that is often deficient in both the users (students at all levels and good number of professors) and the library personnel. In terms of helping the clientele, it seems apparent that consultation services must be collaborative efforts involving both the computer service personnel and the professors and researchers. Among them are specialists in computer science and statistics. But, if one believes that users in search of numerical information are better served in a library (and do not forget that these are DLI participating libraries), it appears problematic to me to

**Table 2. Numerical data use in the context of DLI.**

CD-ROMS delivered to participating universities to DLI	Files downloaded from the FTP site of Statistics Canada		
	1995	1996	1997
1035	887	10173	24384

We have discussed institutional participation, but what happens to the data use level?

In March 1998, more than 1,000 CD-ROMs had been delivered to participating universities. The number and growth of FTP file transfers rose from 10,000 in 1996 to 25,000 in 1997. Obviously, these transfers were not carried out exclusively on data files, but also on command files and text files (e.g., code books, readme files, etc.). Also, transfers do not focus only on micro data files. Aggregated census data accounts for a high proportion of transactions.

Looking at the previous data about DLI, one can advance some observations:

1. Perhaps some data would not have been ordered due to cost, even by libraries that already have data use experience;
2. One can suppose that the global cost would have been far more important if all these data had been acquired individually;
3. It is probable that several files have been downloaded or that CD-ROM products have been ordered by libraries that, until now, had little used numerical data.

Therefore, it seems that DLI contributes largely to data distribution and that the program replies to a real need. The objective to make numerical data accessible at a reasonable price is, therefore, partially attained. On the other hand, it is important to remember that DLI was not conceived merely to reduce the cost to those already using data. The real objective is to increase data use by the whole community, to see a real expertise developed in data use, and to enable a greater number of studies that focus on



think that data information specialists must always cater to other professionals. Collaboration is essential, but total dependence should be limited. Training needs are an important, yet considerable cost of rendering data more accessible.

### **Training**

To satisfy their training needs, DLI representatives can count on a continuous training program put in place by the DLI External Advisory Committee. In addition to this national program, local organisations (such as the Council of Prairie and Pacific University Libraries' (COPUL) Consortium of Library Electronic Data Services (ACCOLEDS), or the Working Group on Data of the Conference of Rectors and Principals of Quebec Universities) also organise training activities.

Since the inception of DLI, several training activities have already taken place. At the initiative of the advisory board, 4 DLI initiation workshops (in fact, the same workshop repeated 4 times) have taken place in four Canadian cities during 1997. These workshops brought together 120 individuals who until then knew almost nothing about numerical data. In evaluating the workshop, participants stated their preference for the organisation of additional workshops on more specific themes.

As a result, new workshops will take place this spring on the use of SPSS for data processing and numerical analysis. In fact, one of these workshops was held two weeks ago in Montreal.

Conducting training sessions raises both challenges and opportunities for librarians interested in promoting numerical data use. It is a challenge because it is necessary to develop a certain level of competence before being able to teach. But challenge aside, the possibility of teaching users represents a privileged opportunity to assume our role as information specialists. Instead of waiting to be asked for information, we can create a demand for information. How can users ask to use numerical data if they do not know it exists or if they can not use it?

Far from me to suggest that librarians replace other professionals or professors. However, I believe that a solid knowledge of data and of the tools of exploitation - from which arises a need for training - is necessary. First, it allows us to exercise the educational aspect of our job and, second, it enables us to become knowledgeable spokespersons alongside professors and other professionals. It is only with a solid knowledge of data that we can become counsellors for users, orienting them towards the best data sources or advising them on the best manner of exploitation. The competence of data librarians is as necessary to establishing bonds with professors and researchers, as is identifying the main research areas for which data are required.

There certainly is not unanimity among colleagues on the way in which these new responsibilities will be handled nor, consequently, on the usefulness of advanced training. Some will never be able (nor want) to develop statistical expertise. The problem is complex and there is certainly no correct reply, but it will have to be considered and a great deal of progress will have to occur.

No matter what others decide to do, some libraries, such as the Carleton University Data Centre, have already specialised in numerical data and offer complete data services that include computer and statistical assistance. Such levels of service are not widely found. Even large universities do not always offer complete data services. All the new libraries that now play a role in data specialisation have not and will not develop such an expertise. But is there a middle road and where is it situated? If data accessibility is essential to democracy, the inequality of services offered constitutes an important obstacle to exercising our rights.

### **Equality of access**

Within the DLI framework, the choice has been made to insure numerical data development in libraries. Considering information needs in general, researchers in small or regional universities are no longer penalised with regard to information accessibility. With the development of new technologies, such as the Internet and the emergence of periodicals and other electronic publications, the availability of large bibliographical data bases on either CD-ROM or the Internet (UNCOVER) and finally, with the development of increasingly specialised inter-library loan services (Ariel), the disparities have lessened between the large and small universities, between urban universities and those in the regions. This is true even if the level of document availability remains variable at the local level.

The situation with regard to numerical data availability is entirely different. It is obvious that students at universities that offer well structured data services are in a favoured position over their colleagues who do not have such access. The other libraries simply do not offer the students the same level of support.

In fact, one can assume that regional disparities were even greater before the inception of DLI and that the program will attenuate these differences. In this regard, the question of training for librarians and their perceived role in offering these new services is an important consideration.

### **Regional disparities**

Another aspect related to the disparity of services offered has to do with the regionalisation of the data. Regional universities are often located in areas that are characterised by a low population density. Research work related to regional problems require data over geographical areas that are not comparable to data that defines large metropolitan

regions. This is the problem experienced at the University of Quebec at Rimouski, which has Masters and Ph.D. programs in regional development. The users from these groups need not just numerical data, but numerical data at a specific geographical level. Unfortunately, the data available to these researchers are often over too wide a geographical level. Even if more specific data exists, they are not available for their use. This is the problem, for example, with the large Survey of Consumer Finances and with the National Population Health Survey.

If the objective of DLI is to increase data access in all universities no matter where they are located, the question of geographical data specificity is important, even though we recognise that solutions will not come directly from DLI. Indeed, the mandate of DLI is to give access to standard data products provided by Statistics Canada. The notion of standard products is also linked to the question of data confidentiality. As a citizen, one can only rejoice in observing that Statistics Canada respects norms of strictest confidentiality and that these principles should never be challenged. From the academic viewpoint, the problem is not less important. On the other hand, it is probable that the successes of DLI will increase demand for this level of data. The question of confidentiality is also important from the standpoint of the installation of data services. It is not sufficient to simply initiate users to the data and to analytical instruments, such as SAS or SPSS, if one cannot also provide data that satisfies their research needs. There is a risk of losing hard-gained credibility if users do not have access to data that they know exist, especially after

they invest considerable energy in learning complex instruments.

### **Conclusion**

Previous statistics have shown that DLI has had real success; a success that exceeds the hopes of many. But beyond all the figures relating to file transactions and considering the current context of data use described here, one of major successes of DLI has been to enlarge data access to a greater number of individuals interested in using numerical data in Canadian libraries. These people share considerable amounts of information (via the two list servers that help in administration of the program). Several of them have had the opportunity to meet during workshops. The community is, therefore, in the process of widening and there exists a tangible willingness among colleagues who are more experienced to share their expertise.

Finally, I would like to mention that universities in Quebec are full participants in this process. All participate in DLI. Members of the sub-workgroup on numerical data files are presently working on the development of an identification and data extraction system that will facilitate and stimulate data use.

\*Paper presented at the IASSIST Conference, May 21, 1998, at Yale University, New Haven, Connecticut. Richard Boily, Université du Québec à Rimouski, and member of the Conference of Rectors and Principals of Quebec Universities (CREPUQ) Working Group on Data



# The California Digital Library: Implications for Social Science Data Files Collections

The recent establishment of the California Digital Library creates an unprecedented opportunity to bring social science data to users in a more user-friendly format as well as making it available to a much wider audience.

The California Digital Library was established at the University of California in the middle of last fall, in effect creating a tenth campus library, this one entirely virtual. In practice, it is based at the UC Office of the President headquarters in Oakland, California, it's main manifestation in its initial year appears to have been acquiring licensing agreements with computerized databases of academic journals in the fields of science and technology. But with an anticipated infusion of \$3 million from state coffers this coming year, plus another \$1 million from the University of California itself, the CDL, as it is called, will soon be looking for new disciplines to cover, and new territory to conquer. According to press reports, California has had a budget surplus of \$4.4 billion which can hopefully be partially allotted to the libraries.

Social Sciences are a likely future field of attention for the CDL, and UC's data archivists and librarians recently met with the CDL's newly installed Collections Officer to discuss matters of mutual interest, including ways of collaborating in possible future joint endeavors.

This paper, then, is part of a continuing, fairly new process of rethinking how the UC system collects and provides services to social science material in digitized formats.

Since the CDL bills itself as the 10th campus library collecting everything in digitized form, there might have been initial fears that collecting of digitized materials would end on the individual campuses that make up the UC system. While legally and constitutionally the University of California is one state-wide system, in actual practice, there is almost no collection development (in terms of library resources) at the system level, except for licenses to databases negotiated system-wide and made available via Melvyl, the UC catalog and database gateway; and large purchases made in the past through "shared purchases" system-wide, largely made up of big ticket items and large microform sets. Each campus is strongly

by Daniel C. Tsang\*

independent, each headed by its own chancellor and strong faculty senates; campus libraries are no different, each with its own collecting focus, depending in a large part on the research and instructional needs of the faculty on the particular campus.

Despite being three years in the planning stage, the actual creation of the CDL may have caught many campus librarians by surprise. Any initial fears that collecting of digitized materials on campus would be displaced by the CDL collecting the same stuff for all campuses soon gave way to the realization that there was no way individual campus digitized collections would disappear, nor would collection development cease. In fact, since data archives and data collection development vary across campuses, meeting individual needs not necessarily duplicated elsewhere, the likely model that will prevail is one more of collaboration with the CDL than one of the CDL usurping and displacing campus collection patterns. While it is relatively easy (and familiar) for the CDL to negotiate licenses over journals in the social sciences, it is a different kettle of fish for the CDL to enter the field of collecting (and making available) raw data that are the products of social science research. It is not just a matter of licensing, but also one of making it available to users — and which users need to be defined — in an appropriate format.

How the CDL is structuring its own collection development may be instructive here. Currently, within its science and technology collection focus, the CDL has selected biotechnology and computer science as the two areas where acquisition of digitized information in formats other than electronic journals or monographs would take place. This is described in a March 8, 1998 announcement on its listserv (CDLINFO-L) as providing a "laboratory for learning and planning for future CDL collections," and for "gradually [adding] other types of content" beyond electronic journals or monographs. This is an ongoing experiment, and it is too early to report any results. But what is clear is that the CDL intends to acquire datasets in biotechnology and computer science initially, although what it would acquire in particular is not apparent.

In addition, the CDL plans to structure its collection in

three “tiers”, namely, Tier 1, material funded, in whole or in part, by the CDL; Tier 2, material funded by two or more campuses; and Tier 3, material that is paid for by only one campus. For the last tier, this implies the CDL would provide a gateway, on its Web site apparently, to material on that campus, even if, as is likely, only users affiliated with that campus can use the material. For tier 2, then, material would likely not be accessible to those outside the campuses that funded it; and for tier 1, since the CDL funded the material, it would presumably be made available to most or all campuses (a campus can opt out of a particular acquisition).

Its current collection principles are that priority should be given to digital format acquisition of those resources which offer economies of scale by benefiting the most faculty and/or students both locally and system wide. Also, electronic materials should be selected based on increase of access to the installed base of UC library collections and build on the investments already made by the university in digital resources. If and when the CDL enters the field of social sciences collecting and social science data in particular, a major issue appears to me to be the one of defining who would be the potential user population. On a traditional campus archive or data library, UC data archivists and librarians have generally acquired material only for the use of faculty and students on that campus. In addition, as I argued in an earlier IASSIST conference paper<sup>1</sup>, collection development, in UC’s as elsewhere, has generally, been reactive rather than proactive; i.e., datasets are acquired in response to specific requests, rather than, as with books, collected in advance of specific request, and often without anticipation of actual immediate use (e.g., through book approval plans). Since that paper, however, with the advent of the main source of social science data now distributing, in effect, its entire newly acquired collection every quarter or so in the form of Periodic Release CD-ROMs, libraries are acquiring data without any selectivity, in actual practice, and certainly not in response to any specific request. In addition, with the ICPSR’s entire archive much easier to retrieve (with the demise of distribution on round tapes), data archivists can potentially replicate much of the collection on their own campuses, although as of yet, there is no mirror site to the ICPSR archive yet established. Such a mirror site might be one area the CDL might fruitfully consider, especially since its mission appears not only to serve the campus-based academic community but also the community at large, although it is still unclear how that would work in practice.

The CDL, after all, is called the California Digital Library, not the University of California Digital Library, strongly implying it is collecting digital material to serve the entire state, i.e., all the people of the state of California. That was the quid pro quo that apparently was necessary for the state legislature to fund the CDL. In announcing the creation of the CDL last fall, UC President Richard C. Atkinson spoke

of creating “UC’s library without walls.” It would be a library allowing “scholars of all ages and interests to range worldwide in their quest for knowledge, using the Internet, the World Wide Web and a computer.”

If the target population is the scholarly community within UC and beyond, that would be one thing. Libraries are used to collecting for scholars; the CDL could very well mirror ICPSR’s archive (after becoming a full-fledged member like the other UC’s that are members) by paying them enough money so that they would not think that they are losing money. But since there is no physical campus associated with the CDL, what does this mean? Or is the CDL membership to take the place of the individual campus memberships? That does not seem likely, given that existing campus archives and data collections have constituencies they have nurtured and served for years; they are not likely to disappear, at least not without a fight. But if the CDL promised access to all ICPSR data, and provided front-end interfaces that facilitated the extraction of variable-level data (it would need to write the software etc.) of selected datasets, would that not make local service points less necessary, or even impractical? The countervailing argument, of course, is that with all secondary analysis of data, it is not sufficient merely to make the data available; there has to be a variety of services (metadata access; interpretation of metadata; statistical consulting; etc.) that, because computer setups vary across campuses, and even within, are best handled at the local level.

In addition, the CDL can perhaps better negotiate licenses at the system-level with data vendors that provide data, such as economic time series, of interest across the UC system.

But given its mandate to serve the state, the CDL and its pioneering vision, the CDL could very well do more than just provide access, even improved access, to what currently exists. The CDL could well get involved in one or more large-scale digitization projects, funded by industry and government, to archive and make accessible datasets previously not readily available, such as government data at both the state, county and municipal levels. The challenge here is for joint partnerships with communities local and state; what is sorely needed is a collaborative effort to make sure that government data does not go the way of the main frame and that they are archived and preserved, and eventually made accessible to users.

As the CDL develops, one potentially controversial area involves intellectual property rights. Who owns the research that faculty have invested their time in? The University is now arguing that it does, and that faculty who sign away rights to articles, for example, are just making commercial journal publishers much more rich. Administrators are now wondering why a university should

have to pay exorbitant fees to access the journal output of their own faculty, just because a commercial journal published it. Richard Lucier, the CDL's head, argues that scholarly publishing must change, and that universities must on their own, compete with the industry and "publish" on line the scholarly output.

With data however, is it likely that individual faculty, even at UC, will be willing to deposit their research data at the CDL, if the University insists that it must? The University recently revised its policy on research; now, even if faculty use a dataset gathered elsewhere for secondary analysis, it must register with local Institutional Review Boards. Local boards could well insist that faculty deposit data thus gathered (or originally collected data for that matter) with the CDL, or the campus data archive. Right now there is no such provision or mandate, not least because researchers are unlikely to be willing to part with their data, and there is no common understanding of who owns what. If a faculty member leaves, he or she is allowed to take research data he or she has gathered; thus far, I am not aware that the university has insisted on ownership. But a case could well be made, given that every employee is, upon hire, made to sign away most of his or her patent rights to the University.

Involvement in large-scale projects is especially likely, and necessary, if indeed, the user community stretches beyond the scholarly community. If indeed the vision is to let anyone access the information (and since the CDL is called the California Digital Library, not the University of California Digital Library, as originally envisioned), that suggests that anyone, "of any age", would have access to

the library. That would well be a mammoth task, devising a dataset (or more) that would be useful to such a mythical user.

There are many more issues one could raise, not least that of digital archive maintenance, authentication, and dataset updating, as well as version control. At a minimum, the CDL could provide a union list of what exists in existing campus archives and collections, providing some bibliographic control to an existing situation that is anarchistic at best. But I see it as doing more than that; it can best make the use of data more appealing, by providing the necessary tools to access data from the hard-to-find to those most popularly requested. As such the CDL can help those of us in the data archive community by making, and educating, more people to be sophisticated data users and consumers.

In conclusion, the CDL is unlikely to be the sole digital repository for California of social science data, but its creation and expansion will likely spur collaborative efforts with existing collections and archives as well as create new ways to provide improved access to these collections.

1. Daniel C. Tsang, "Academic Libraries and Collection Development of Nonbibliographic Machine Readable Data Files," *IASSIST Quarterly*, volume 12, number 3, Fall 1988, pp. 47-55.

\* Paper presented at the IASSIST Conference, May 21, 1998, at Yale University, New Haven, Connecticut.



---

# Privacy and Computerized Data Bases

In the United States, privacy legislation generally has been limited to government records at the different levels of government, i.e., Federal, state and local. These laws impose requirements and restrictions on how government agencies collect, maintain and use information about individual persons. The United States, with few exceptions, has not adopted legislation to regulate and restrict the collection, maintenance or use of personally identifiable information by non-governmental entities. In other countries, this type of legislation is frequently referred to as "data protection laws." However, if a threat to the personal privacy exists today in the United States, it comes not from the regulated governmental data bases but from the non-regulated ones outside of government buildings.

---

*by Thomas E. Brown \**

---

The information in these data bases has long be available in paper form without threatening individual privacy. But technological developments have altered the situation. First, with the spread of technology, data is increasingly available in electronic form. Second, computer processing speeds have increased. With data mining tools, a data base query that took six minutes in 1994 now takes less than nineteen seconds. Third, with the availability of increased computing speed, it is now easier to combine data from multiple sources and create comprehensive information products. Fourth, the cost of storing electronic information — on-line, near-line and off-line storage — has dropped. Fifth, personal computers are becoming more and more affordable and thus more wide spread. Finally, with the spread of the personal computers, Internet use is becoming commonplace from businesses and homes.<sup>1</sup>

Technology, then, has permitted the growth of what Vice President Gore has called "profiling," or the ability to build dossiers about individuals by aggregating information from a variety of database sources. These dossiers now have detailed information on the vast majority of the American population, including children. For example, Axiom Corporation has information on 196 million Americans on 700,000 data tapes with 350 trillion bytes. Information America claims that it has employment and other demographic information on 160 million individuals, 92 million households, 71 million telephone numbers, and 40 million deceased persons. Finally, Medical Marketing

Service, Inc., offers its "Patient Direct" data base with information on more than 20 million households documenting an individual's age, gender, income, educational status, and health condition, from allergies (4.3 million households) to yeast infections (1.4 million)<sup>2</sup>.

The sources of information are varied. First, governmental records provide a wealth of information on individuals and, under a variety of Freedom of Information laws, are readily available. While some invasive records are from the Federal Government, most are seemingly from state and local governments. These records concern real estate transactions and holdings, marriage and divorce, birth certificates, driving records, drivers' licenses, vehicle registrations, civil and criminal court proceedings, paroles, postal service change-of-address forms, voter registrations, bankruptcies and liens, incorporation documents, workers' compensation claims, political contributions, firearm permits, occupational and recreation licenses, and filings with the Securities and Exchange commission. For example, the Federal Aviation Administration has a list of all individuals licensed to fly in the United States which includes certification class, medical certificate and date of last medical examination. Real estate documents describe the property, dates of sales, selling prices, mortgage amounts, lender, and the names of the sellers and purchasers. Social security numbers are readily available as well, most commonly from state departments of motor vehicles, which also provide individuals' name, address, height, weight, gender, eye color, date of birth, and whether or not an organ donor. For example, New York's publicly available drivers' information includes vehicle ownership, accident reports, conviction certificates, police reports, complaints, hearing records, suspension and revocation orders. Although government records are increasingly available in electronic form and electronic FOIA guarantees access to the records in that format, other records must first be digitized.<sup>3</sup>

Non-governmental entities, but still publicly available information, offer additional sources of information. For example, newspaper and magazines identify and provide background information on individuals, and electronic editions are now frequently available from the Internet. Powerful search tools permit people to search these



newspapers and magazines and find all references to a given individual. Many professional and alumni organizations make membership lists available on Web sites. Indeed, many Web sites contain detailed personal information to anyone who clicks on the URL, such as adoption pages where adopted children and birth parents post detailed personal information in hopes of connecting with their blood relatives.<sup>4</sup>

A third type is private proprietary information. Maybe the most well known of this type is the information of the three major credit reporting agencies, Trans Union, Equifax, and Experian. But other even more invasive data bases exist with varying degrees of confidentiality associated with them. More than 3,000 corporations in the United States collect, maintain and sell information from their data bases for marketing purposes. Most of the information is provided by the individuals themselves, often through warranty cards or product registration cards. A careful reading of these forms reveals the depth of personal information people supply. As an example, a recent card for a coffee bean grinder asked about sex and ages of all household members, marital status, occupation, income, educational level, credit cards, home ownership, anticipated changes during the next six or 12 months, and finally a list of sixty interests activities in which the respondent participates in regularly. Another series of incredibly revealing data bases are those maintained by banks and credit card companies which list the individual transactions charged to each credit card. But far more sinister is the spread of supermarket customer cards. Last year, a survey indicated that 60 per cent of the supermarkets in the country had started such a program or intended to do so soon. These cards are used to record in a data base each product which a consumer purchases. One's shopping habits reveal a lot about that individual. The purchases of condoms or large quantities of alcoholic beverages connote a life style. Patent medicines reveal aspects about one's health, and purchases of certain brands and products betray one's value systems and interests. Every time somebody makes a telephone call, it creates a record in a database which offers much information to the phone company and other marketers. Through billing records, local carriers and long-distance carriers learn whom people are calling and when and where calls are made. Finally, the Internet itself is a great collector of personally revealing information. Commercial Web sites collect personal information through a variety of means, including registration forms, user satisfaction surveys, contests, and order forms. For example, an online doctor-referral service asks users for their name, mailing address, e-mail address, insurance company, and whether they want information on a variety of health concerns, such as urinary incontinence, hypertension, cholesterol, prostate cancer, or depression. Another Web site is from a mortgage company for pre-qualification for home mortgages. Potential borrowers provide their names, social security numbers, home and

work telephone numbers, e-mail addresses, previous addresses, current and former employers, lengths of employment, income, savings, and credit histories including credit cards. In the Spring of 1998, a quarter-million people completed a detailed survey at the ESPN Web site to enter a lottery for tickets to the NCC Final Four tournament. Even without telling the users, Web vendors can collect detailed personal information as they can identify which pages the user visited, what the user bought, where the user linked from and where the user went on the "Net" when he or she left the site, and in some cases, how long an individual was at the site and on each page of the site. Recently reported, some of the largest commercial sites including Lycos-Tripod and Geocities, had begun providing users' reading, shopping, and entertainment habits to a centralized system which links that information to the user's age, income, ZIP code, number of children, and information obtained from on-line forms. To protect privacy, the system uses a unique identifying number associated with the hard drive of the computer accessing the sites. Thus when the user connects to a participating site which recognizes the number on the drive, targeted sales messages will be sent. But already, entrepreneurs are trying to link the preferences in compiled computerized data bases with mailing lists of traditional direct marketers. The owners of the system have announced that it will not collect information about sexual interests or health related topics. But with money on the line, such a voluntary restriction is probably not universal or permanent. For example, some Internet sites devoted to specific diseases have solicited data from site visitors and then sold that information, either directly or indirectly through data intermediaries, to companies marketing drugs or other therapies for the specific disease. As Nat Goldhaber, chair of a Web vendor called Cybergold, commented, "What the Internet has done is make explicit what used to be implicit — namely that there dossiers on you than can be built up with great granularity."<sup>5</sup>

As June 24, 1997, there were fifty-one database vendors and information bureaus which will sell detailed personal information on the Internet about telephone use, assets, criminal histories, vehicle and driving records, aircraft, boat and gun ownership and usage, business materials, marriages and divorces, current and previous addresses, and information on neighbors and relatives. Some have unexceptional names, such as Discreet Research whose slogan is "When you need to know." Others are more interesting, such as Dig Dirt, Inc., whose saying is "Because what you don't know does hurt you."<sup>6</sup>

In the collection, maintenance, and use of this information, the United States has adopted as essentially *laissez faire* approach. Current American privacy law has been described as "sectoral," that is "a handful of disparate statutes directed at specific industries that collect personal data." In fact, the Federal level has only six data protection



laws in place. The first and clearest example is the Fair Credit Reporting Act which guarantees the individual the right to gain access to personal information which the credit reporting agency has and the right to dispute erroneous information and add corrective details. It also generally restricts access to those businesses to which the consumer has applied for credit, insurance, employment, or a lease agreement. The second is Federal Educational Records Privacy Act (FERPA) or more commonly known as the Buckley Amendment. This limits the release of students' educational records to educational personnel and educational institutions. Two other acts are the Cable Communications Policy Act of 1984 which restricts cable television subscriber information and the Telecommunications Act of 1996 which governs customer proprietary network information. The last data protection law in the United States restricts the release of an individual's video tape rentals. Enacted as a reaction to publicizing Judge Bork's video tape rental during his Supreme Court confirmation hearings, the law prompted Secretary of Health and Human Services Donna Shalala to comment, "Our private health information is being shared, collected, analyzed and stored with fewer Federal safeguards than our video store records." In the absence of legislation, case law is working against data protection. In *United States v. Miller* [425 U.S. 435 (1976)], the Supreme Court ruled that individuals have no constitutional protection of information which that they have voluntarily provided.<sup>7</sup>

But currently, some 80 bills to strengthen the data protection are pending in the Congress. Some would restrict the dissemination or use of the social security numbers; others would allow individuals to stop the post office from selling their change of address requests; and at least one would establish an independent regulatory commission to control the collection, maintenance, and use of individually identifiable information. According to one observer, the *New York Times*' Nina Bernstein, "But with few exceptions, the proposals seem to be going nowhere. Beneath the surface of their popular appeal, most are mired in unresolved conflicts over contradictory goals: on the one hand, preserving personal privacy, and on the other, the advantages of quick, computerized access to personal information for fighting crime, fraud and waste, or promoting growth in the information economy."<sup>8</sup>

The one exception will be data protection of health information. The Health Insurance and Portability and Accountability Act of 1996, commonly known as Kennedy-Kassebaum Act, required the Clinton Administration to propose legislation on the creation, maintenance and use of medical information on individuals by September 30, 1997. [Ironically, this same legislation also required that the Administration create a standard medical identifier so individuals' medical records could be widely available, regardless of the health care program, from a database of

health information.] If the Congress does not enact legislation by August 1999, then the Administration must issue regulations. So on September 11, 1997, Secretary Shalala proposed legislation. But the Administration's proposal is only one of several laying in the legislative hopper. Reaching a consensus will not be easy in the tug-of-war between consumer groups, law enforcement agencies, and health care professionals. Should law enforcement officials need a warrant or court order to get medical records? Can records obtained investigating insurance fraud be used for other criminal prosecutions, such as information about illegal drug use lead to prosecution? If consumers have the right to change or delete medical information, then some argue that this could endanger a vital resource needed for medical research, for public health analyses, and for improved medical care.<sup>9</sup>

If a legislative consensus is difficult, self-regulation is an option which conforms to the *laissez faire* approach and has an honored tradition in the country's history. Last December, the Federal Trade Commission announced an agreement amounting to self-policing by "individual reference services," that are businesses which have been selling personal information to the general public. By the end of this year, fourteen of the largest of these organizations, including the three major credit reporting agencies and the largest direct mail marketers, announced that they would no longer provide information to the general public. They would also limit the types of personal information they would provide to commercial users like marketers, banks, lawyers and journalists. Yet they would still allow unrestricted access by law enforcement personnel, licensed private investigators, and corporate security staff. According to *The New York Times*, this agreement embodied the Administration's strategy of self-regulation. "The agreement sets in motion the first meaningful trial of the Clinton Administration's privacy policy, the stated goal of which is to protect individual privacy in the Internet age without resorting to new laws and regulations." Privacy proponents have objected to the agreement as too little. While individuals can request that their records be erased from some data bases, they cannot access all of the information being maintained and disseminated about them, and to have themselves removed from selected data bases, individuals would have to contact each of the fourteen reference services separately. An obvious loophole will be for an individual to hire an attorney or private detective to serve as an intermediary.<sup>10</sup>

In this same vein of self-regulation, the 3000-member Direct Marketing Association has issued "Guidelines for Personal Information Protection" which stipulates that personal data collected for marketing purposes should be only for that purpose. It further maintains to its Committee on Ethical Business Practices to investigate the misuse of marketing information. The Association has also announced that it will require, beginning next year, its

members to disclose publicly how they gather and use data. The Council of Better Business Bureaus is working on a model for self-regulation that would impose sanctions on businesses that fail to follow a code of conduct to protect people's privacy. A recent innovation to self-regulation is incorporating a seal onto a commercial Web site to indicate that the site follows an established code of conduct regarding privacy. A nonprofit group called "TRUSTe" already has a system in place for about 120 companies. In July, another group, Online Privacy Alliance, emerged with the same intent.<sup>11</sup>

These efforts at self-regulation have been haphazard at best. The Federal Trade Commission tersely concluded earlier this summer, "To date, . . . the Commission has not seen an effective self-regulatory system emerge." The Clinton Administration began to hedge last May when Vice President Gore outlined a new administration initiative on privacy. It consisted of a renewed call for legislation regarding medical records, a Federal Web site to assist consumers in deleting their names from commercial data bases, creation of the privacy officer in every Federal agency, and a conference to address the topic in June 1998. This speech, calling for an "Electronic Bill of Rights," was interpreted as an admission that self-regulation had not been as effective as the Administration had hoped. According to Janlori Goldman, a noted privacy specialist at Georgetown University, "This Administration has been singing the praises of self-regulation for some time now, but this is an acknowledgment that there are significant limits to what the private sector will do on its own." As far as the legislation concerning medical records is concerned, Dr. Goldman opined that the Clinton Administration is "using medical privacy as a signal to the public and a stick to industry to say that we have a history of abuse in this area and the Administration wants to do something about it."<sup>12</sup>

But regardless of the mode of data protection — legislation, regulation or self-policing, discussions of data protection revolve around eight principles of fairness: openness, individual participation, collection limitation, data quality, use limitation, disclosure limitation, security, and accountability. While originally proposed by a United States federal advisory committee, the principles were most clearly codified in the Council of Europe's *Convention for the Protection of Individuals with Regard to Automatic Processing of Personal Data*. For the United States to be full economic partners with Europe, data protection efforts will have to conform to the Council of Europe's *Convention*. Two statements in the *Convention* should give pause to archivists, namely:

"Article 5 - Quality of data

Personal data undergoing automatic processing shall be:  
b. stored for specified and legitimate purposes and not

used in a way incompatible

with those purposes; . . .

e. preserved in a form which permits identification of the data subjects for no longer

than is required for the purpose for which those data are stored."<sup>13</sup>

This concept that personal data can be used only for the reason for which it was collected is fairly common in the discussions of data protection. In the above discussion of the Direct Marketing Association, its "Guidelines for Personal Information Protection" stated that personal data collected for marketing purposes should be only for that purpose. In discussing the Clinton Administration's position on health care, Secretary Shalala stated that personal medical information should be "for health care and health care only" with very few exceptions. But from an archival perspective, this limitation on use has a potential problem. Archival theory discusses the primary value of records as the purpose for which the records were created. This is in contrast to the secondary value of records which is the value of the records to someone other than the record's creator for a reason other than that for which they were created, and it is the secondary value of records that justifies the archival retention of records after the record's creator no longer needs them in the course of business, but if the use of personal data is limited to the reason for which it was collected and if the archives are not exempt from this limitation, then records cannot be retained in archives for their secondary values. Indeed, under the Council of Europe's *Convention*, the records must be destroyed as soon as their primary value has ended. In terms of medical records, for example, if an exception is not made for archival retention of some personal medical information then the history of medicine, the history of technology and the history of science — all currently viable fields of historical study — will be severely curtailed. Application of the *Convention's* beyond just medical records will threaten the continued existence of records important to future genealogists and family historians. Obviously, a solution is to outline exceptions to the use of limitation provisions in any data protection effort — whether legislation, regulation, or self-policing. One of these exceptions must be for the archival retention of records with significant secondary values to be released only after the passage of time that would permit access without endangering the privacy of individuals. Unfortunately, the possibility of records having secondary values seldom enters into the debates over data protection, but then it seems that the archival profession has seldom entered in the debates over data protection.<sup>14</sup>

<sup>1</sup>Federal Trade Commission, *Individual Reference Services: A Report to Congress*, December 1997, pp. 3-4.

<sup>2</sup> "Vice President Gore Announces New Steps Toward an Electronic Bill of Rights," This Week's Press Briefings and Releases, July 31, 1998, <http://library.whitehouse.gov>, August 3, 1998; Robert O'Harrow, Jr., "Data Firms Getting Too Personal?" *The Washington Post*, March 8, 1998, pp. A1, A18-A19; Federal Trade Commission, *Individual Reference Services*, pp. 36; Sheryl Gay Stolberg, "The Numbering of America: Medical I.D.'s and Privacy (Or What's Left of It)," *The New York Times*, July 26, 1998, p. WK3.

<sup>3</sup>Federal Trade Commission, *Individual Reference Services*, pp. 4-6, 37; Nina Bernstein, "High-Tech Sleuths Find Private Facts Online," *The New York Times*, September 15, 1997, electronic edition.

<sup>4</sup>Federal Trade Commission, *Individual Reference Services*, p. 5.

<sup>1</sup>Federal Trade Commission, *Individual Reference Services*, p. 3; O'Harrow, March 8, 1998, p. A18; Nina Bernstein, "Lives on File: Privacy Devalued in Information Economy," *The New York Times*, June 12, 1997, electronic edition; Federal Trade Commission, *Privacy Online: A Report to Congress*, June 1998, p.3, 39; Saul Hansell, "Big Web Sites to Track Steps of Their Users," *The New York Times*, August 16, 1998, pp. 1, 24; Denise Caruso, "Who Knows What About Whom on the Internet," *The New York Times*, April 13, 1998, electronic edition

<sup>5</sup>Federal Trade Commission, *Individual Reference Services*, p. 3; O'Harrow, March 8, 1998, p. A18; Nina Bernstein, "Lives on File: Privacy Devalued in Information Economy," *The New York Times*, June 12, 1997, electronic edition; Federal Trade Commission, *Privacy Online: A Report to Congress*, June 1998, p.3, 39; Saul Hansell, "Big Web Sites to Track Steps of Their Users," *The New York Times*, August 16, 1998, pp. 1, 24; Denise Caruso, "Who Knows What About Whom on the Internet," *The New York Times*, April 13, 1998, electronic edition.

<sup>6</sup>See <<http://www.dresearch.com/>> and <<http://www.pimall.com/digdirt/moore.htm>>.

<sup>7</sup>Federal Trade Commission, *Privacy Online*, p. 62; 15 USC 1681; 20 USC 1232g; 47 USC 551; 47 USC 222; 18 USC 2710; Robert Pear, "Clinton to Back a Law on Patient Privacy," *The New York Times*, August 10, 1997, p. 22.

<sup>8</sup>Peter Maas, "How Private Is Your Life?" *Parade Magazine*, April 19, 1998, p. 5; Nina Bernstein, "Goal Clash in Shielding Privacy," *The New York Times*, October 20, 1997, p. A16.

<sup>9</sup>Steven Findlay, "Prescription for Patient Privacy: Administration today offers plan to ensure confidentiality," *USA Today*, September 11, 1997, pp. 1-2; Editorial, "HHS identifier puts privacy at risk," *Federal Computer Week*, July 20, 1998, p. 24; Arthur Allen, "Exposed," *The Washington Post Magazine*, February 8, 1998, pp. 11-15, 27-28.

<sup>10</sup>Federal Trade Commission, *Individual Reference Services*, passim; Katherine Q. Seelye, "A Plan for Database Privacy, But Public Has to Ask for It," *The New York Times*, December 18, 1997, pp. A1, A24; John Markoff, "Guidelines Don't End Debate on Internet Privacy," *The New York Times*, December 18, 1997, pp. A24.

<sup>11</sup>Federal Trade Commission, *Individual Reference Services*, p. 38; O'Harrow, March 8, 1998, p. A18; Robert O'Harrow, Jr., "White House Effort Addresses Privacy," *The Washington Post*, May 14, 1998, p. E1, E4; Robert O'Harrow, Jr., "Internet Companies Move to Safeguard Computer Users' Privacy," *The Washington Post*, July 22, 1998, p. A13.

<sup>12</sup>Federal Trade Commission, *Privacy Online*, p. 41; John M. Broder, "Gore to Announce 'Electronic Bill of Rights' Aimed at Privacy," *The New York Times*, May 14, 1998, p. A22; O'Harrow, May 14, 1998, p. E4.

<sup>13</sup>Robert Gellam, "Don't fear privacy protection — arm yourself with fairness checks," *Government Computer News*, May 4, 1998, p. 26; Department of Health, Education and Welfare, Secretary's Advisory Committee on Automated Personal Data Systems, Records, Computers, and the Rights of Citizens, (July 1973). Council of Europe, European Treaties, ETS No. 108, Convention for the Protection of Individuals with Regard To Automatic Processing of Personal Data, Strasbourg, 28.I.1981, <http://www.coe.fr/eng/legaltxt/108e.htm>.

<sup>14</sup>Federal Trade Commission, *Individual Reference Services*, p. 38; Pear, p. 22; T. R. Schellenberg, *Modern Archives: Principles and Techniques* (Chicago, University of Chicago Press, 1956), pp. 28-32; T. R. Schellenberg, *The Appraisal of Modern Public Records*, Bulletins of the National Archives, Number 8 (Washington DC: U.S. Government Printing Office, 1956), p. 6.

\* Paper presented at the IASSIST Conference, May 21, 1998, at Yale University, New Haven, Connecticut. Thomas E. Brown, Manager, Archival Services Electronic and Special Media Records Services Division National Archives and Records Administration.



# Data Protection and Privacy in the United States and Europe

## Introduction

The rapid expansion in electronic communications and commerce over the past several years has raised concerns in the United States over personal privacy in an online environment. These concerns have captured the attention of the public, the media, and policy-makers, and there is new interest in the United States in explicit policies protecting the privacy of electronic transactions and personal information. These efforts continue a pattern of policies directed at subject-specific information, such as the *National Education Statistics Act of 1994* that tightened access to personal data collected in the field of education [1].

This pattern is a sharp contrast to the privacy and data protection policies in Europe. Where the U.S. approach has been to provide specific and narrowly applicable legislation, in Europe there are unified supra-national policies for the region. Most countries have implemented these policies with omnibus legislation. The European legislation outlines a set of rights and principle for the treatment of personal data, without regard to whether the data is held in the public or private sector. In the United States, the legal tradition is much more concerned with regulating data collected by the federal government. This paper will review and contrast the development of data protection policies in the United States and Europe.

## What Is Privacy?

Privacy is an important, but illusive concept in law. The right to privacy is acknowledged in several broad-based international agreements. Article 12 of the *Universal Declaration of Human Rights* and Article 17 of the *United Nations International Covenant on Civil and Political Rights* both state that, "No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks." The international concept of "the right to privacy" traces its roots to the U.S. Constitution and to common law [2].

A hallmark article in the *Harvard Law Review* in 1890 is widely credited as establishing the right to privacy as a tradition of common law [3]. In that article, Samuel

---

by Jean Slemmons Stratford &  
Juri Stratford \*

---

Warren and Louis Brandeis defined that right as "the right to be let alone" [4]. They argued that the right to privacy that afforded to intellectual and artistic property in common law is founded, not on principle of protection of private property, but on that of "inviolate personality" [5].

The term "privacy" does not appear in the U.S. Constitution or the Bill of Rights. However, the U.S. Supreme Court has ruled in favor of various privacy interests-deriving the right to privacy from the First, Third, Fourth, Fifth, Ninth, and Fourteenth Amendments to the Constitution. In 1977 in *Whalen v. Roe*, the Supreme Court first recognized the right to information privacy [6]. It noted that the Constitution protected two kinds of individual interests: "One is the individual interest in avoiding disclosure of personal matters, and another is the interest in independence in making certain kinds of important decisions" [7]. Several other decisions have balanced the right to privacy against other compelling interests. The Supreme Court upheld a New York law that required the state to maintain computerized records of prescriptions for certain drugs because the program did not pose "a sufficiently grievous threat" [8]. In *Nixon v. Administrators of General Services*, the Court upheld the federal statute that required national archivists to examine written and recorded information accumulated by the president [9]. The Court ruled that while "the appellant has a legitimate expectation of privacy in his personal communications," that right must be weighed against the important public interest in preservation of materials [10]. The Court did not believe that the appellant's privacy interest was a match for the competing public interest [11].

## U.S. Data Protection Laws

There is no single law in the United States that provides a comprehensive treatment of data protection or privacy issues. In addition to the constitutional interpretations provided by the courts and the international agreements mentioned above, there have been a number of laws and executive orders dealing specifically with the concept of data protection. The most important and broad based of these laws are the *Privacy Act of 1974* and the *Computer Matching and Privacy Act*. These laws deal exclusively with personal information held by the federal government

and do not have any authority over the collection and use of personal information held by other private and public sector entities.

The *Privacy Act* (PL 93-579) is a companion to and extension of the *Freedom of Information Act* (FOIA) of 1966. FOIA was primarily intended to provide access to government information. It did exempt the disclosure of personnel and medical files that would constitute "a clearly unwarranted invasion of personal privacy" [12]. This provision was initially used to deny access to people requesting their own records. So the *Privacy Act* was also adopted both to protect personal information in federal databases and to provide individuals with certain rights over information contained in those databases. The act has been characterized as "the centerpiece of U.S. privacy law affecting government record-keeping" [13]. The act was developed explicitly to address the problems posed by electronic technologies and personal records systems and covers the vast majority of personal records systems maintained by the federal government. The act set forth some basic principles of "fair information practice," and provided individuals with the right of access to information about themselves and the right to challenge the contents of records. It requires that personal information may only be disclosed with the individual's consent or for purposes announced in advance. The act also requires federal agencies to publish an annual list of systems maintained by the agency that contain personal information.

The law had originally proposed the creation of a privacy protection commission; however, then President Gerald Ford was opposed to such a bureaucracy. He wrote

I do not favor establishing a separate Commission or Board bureaucracy empowered to define privacy in its own terms and to second-guess citizens and agencies. I vastly prefer an approach which makes Federal agencies fully and publicly accountable for legally-mandated privacy protections and which gives the individual adequate legal remedies to enforce what he deems to be his own best privacy interests [14].

As a compromise, central oversight was assigned to the Office of Management and Budget, and OMB has exercised relatively weak leadership in the implementation of the *Privacy Act*. The law also calls for the designation of Privacy Act officers within federal executive agencies to handle requests and insure compliance with the code of practice. Ultimately enforcement rests with the courts (as individuals bring suit to redress perceived grievances).

Under the umbrella of the *Privacy Act*, Congress has also enacted the *Computer Matching and Privacy Protection Act of 1988* (PL 100-503). This act amended the *Privacy Act* by adding new provisions regulating the use of computer matching. Computer matching is the

computerized comparison of information about an individual for the purpose of determining eligibility for Federal benefit programs, or for the purpose of recouping payments or delinquent debts under such programs.

In general, matching programs involving Federal records must be conducted under an agreement between the source and recipient agencies. This agreement describes the purpose and procedures for the matching and establishes protections for the matched records. The agreement is subject to review by a Data Integrity Board and each agency involved in matching activities must establish such a board. While the law provides no special access rights to individuals; agencies must notify individuals of any findings based upon a computer matching program before taking any adverse actions; and individuals must be given the opportunity to contest such findings.

The *Computer Security Act of 1987* (PL 100-235) also deals with personal information in federal record systems. It protects the security of sensitive personal information in federal computer systems. The act establishes government-wide standards for computer security and assigns responsibility for those standards to the National Institute of Standards. The law also requires federal agencies to identify systems containing sensitive personal information and to develop security plans for those systems.

### **Narrowly Applicable Laws**

There are also numerous narrowly applicable laws on privacy and data protection. These laws generally fall into two distinct categories. The first governs the status of information held by the federal government. In general, these laws provide declarations regarding the confidentiality of specific types of personal information, provide guidelines for their disclosure and penalties for infringement of the individual's right to privacy.

As examples, 13 U.S.C. 9 absolutely prohibits any use of personally identifiable data from the Census except by sworn officers and employees of the Census Bureau. Similarly 42 U.S.C. 242m protects against the disclosure of personal information gathered by the National Centers for Health Services Research and for Health Statistics for research purposes. The *National Education Statistics Act* (PL 103-382) re-authorized and amended provisions for the National Center for Educational Statistics and the National Assessment of Educational Progress. The act dramatically revised the confidentiality and dissemination practices of the center. The *Tax Reform Act* (PL 94-455) makes tax returns and return information confidential, permits only limited disclosure of returns and returns information for specific purposes, and specifies procedures for disclosure. The law also authorizes persons whose tax returns or return information is disclosed in violation of this Act to bring a civil action for damages and costs of the action, and establishes criminal penalties for wrongful disclosures.



The United States has largely avoided legislation governing the treatment of sensitive personal information in records systems held by sources other than the federal government. The few laws that deal with these systems tend to address the treatment of personal financial information. For example, *The Fair Credit Reporting Act* (90-321) regulates the use of individual personal and financial information by consumer credit reporting agencies. It assures that information is accurate and complete, relevant to the purpose for which it is used, and upholds the individual's right to privacy.

A limited number of laws have been passed to deal with issues outside the financial arena. These laws have generally been implemented in response to specific perceived abuses. As an example, the *Video Privacy Protection Act of 1988* (PL 100-618) amends the Federal criminal code to prohibit, with certain exceptions, the disclosure of video rental records containing personally identifiable information. It permits any person who is aggrieved by a violation of this Act to bring a civil action for damages; and requires the destruction of personally identifiable records within a specified period of time. This law was passed in the wake of criticism following the release and publication of Robert Bork's video rental records, during his consideration as a nominee to the Supreme Court [15].

Several U.S. laws do restrict the federal government's access to records held by other sources. Until the rise of the Internet, misuse of personal data held by entities other than the federal government did not command much attention from policymakers as a threat to privacy or personal liberty. However, government access to these records did seem to be a cause for concern as several laws have restricted federal access to information held in such systems. Typically, agencies must obtain permission or a court order to get access to these records [16].

### Data Protection in Europe

There are two important supra-national policies in Europe in relation to data protection. The first is the Council of Europe's *Convention on Data Protection*, and the second is the EU *Data Directive*. In contrast to U.S. privacy law, privacy protection in Europe is addressed by omnibus legislation covering both public and private sectors

The Council of Europe was set up after the Second World War to help unite Europe by

fostering closer relations between the states belonging to the community, ensuring economic and social progress by common action to eliminate the barriers which divide Europe ... and promoting democracy on the basis of the fundamental rights recognized in the constitutions and laws of the Member States and in the European Convention for the Protection of Human Rights and

Fundamental Freedoms [17].

That *Convention* recognizes the right to privacy as one of the fundamental human rights. The Council's concern with the processing of personal information grew slowly with advances in information technology and the increase in the use of such data. In the late 1960s, the Council's Committee of Experts on Human Rights conducted a survey with regard to human rights and modern scientific and technological developments. It concluded that existing laws did not provide adequate protection for individuals given the developments in these areas. Several other committees examined various aspects of the problem and came to similar conclusions. In 1976, the Council established a Committee of Experts on Data Protection that reported its findings in early 1979 and the result was the Council of Europe's *Convention for the Protection of Individuals with Regard to Automatic Processing of Personal Data*. The Council of Europe *Convention* sets forth the data subject's right to privacy, enumerates a series of basic principals for data, provides for transborder data flows, and calls for mutual assistance between parties to the treaty including the establishment of a consultative committee and a procedure for future amendments to the convention [18].

The Commission of the European Community recommended that member states ratify the Council of Europe *Convention* and warned that it might introduce its own directive on the subject. When it did so, the primary purpose of the directive was to further standardize the level of protection across the Community. The EU *Data Protection Directive* reaffirms the principals outlined in the Council of Europe *Convention* [19].

Major components of the Directive acknowledge the individual's right to privacy. The Directive sets standards for the treatment of personal data collected from individuals and for individuals rights of access, notification, and correction. Of particular interest to the United States is the Directive's treatment of data transfers to countries outside the EU. Article 25 governs the "Transfer of Personal Data to Third Countries." EU Member States may transfer personal data only after determining that "the third country in question ensures an adequate level of [data] protection." The EU shall consider the "rules of law...in the third country" to make this determination.

The Directive was adopted in October 1995, and called for member states to bring their national privacy laws into compliance within three years. These national laws are now going into force across Europe.

The absence of generic privacy legislation in the U.S. is a major concern to the EU nations and this will make determination that the U.S. ensures an adequate level of

protection unlikely. While there are concerted efforts in the administration calling for privacy legislation covering various types of data (e.g. Secretary of Health and Human Services Shalala made recommendations to Congress on the Confidentiality of Individually-Identifiable Health Information on September 11, 1997,) the large number of bills in Congress dealing with privacy issues suggests that the U.S. may continue to take a piece-meal approach to privacy legislation [20].

However, the EU is unlikely to issue an across-the-board finding that U.S. privacy protections are inadequate. The EU could demonstrate its seriousness about the Directive by initially singling out one or more U.S. companies or sectors as not meeting the adequacy test; e.g. any company handling personal medical information [21].

Given policy traditions in the U.S., it is likely that data protection in the private sector will be largely self-regulatory. The Federal Trade Commission has been working with the private sector to develop voluntary codes of conduct, but it is unclear where these efforts will lead. It is difficult to say whether the EU will be able to recognize such an approach as adequate. If the EU decides that the largely self-regulatory approach followed by the U.S. is not sufficient to justify an adequacy finding, a much broader embargo is possible [22]. Privacy and data protection are likely to continue to be big issues in U.S. domestic and international policy. It will be interesting to see how these issues will resolve themselves, or if there is to be a major clash between the U.S. and Europe.

#### Footnotes

[1] *U.S. National Education Statistics Act of 1994*, P.L. 103-382 U.S.C., 9001-9012.

[2] United Nations, General Assembly, 3<sup>rd</sup> Session. "Resolution 217A Universal Declaration of Human Rights," 1948. *The International Covenant on Civil and Political Rights* was adopted by the General Assembly of the United Nations in *Resolution 2200 (XXI)* of 16 December 1966. For the full text of the Resolution and the Covenant, see *Official Records of the General Assembly*, Twenty-first Session, Supplement No. 16 (A/6316), 49.

[3] Samuel D. Warren and Louis D. Brandeis, "The Right to Privacy," *Harvard Law Review* 4 (1890):193-220.

[4] Warren and Brandeis, 193.

[5] Warren and Brandeis, 205.

[6] *Whalen v. Roe*, 429 U.S. Reports (February 22, 1977), 589-604.

[7] *Whalen v. Roe*, 599-600.

[8] *Whalen v. Roe*, 600.

[9] *Nixon v. Administrators of General Services*, 433 U.S.

*Reports* (28 June 1977), 425-484.

[10] *Nixon v. Administrators of General Services*, 465.

[11] *Nixon v. Administrators of General Services*, 465.

[12] *Freedom of Information*, Title 5 U.S.C. 552(b) (6).

[13] Robert Aldrich, "Privacy Protection Law in the United States," (NTIA Report 82-98) in U.S. Congress. House. Committee on Government Operations. *Oversight of the Privacy Act of 1974: Hearings*. 98<sup>th</sup> Congress, 1<sup>st</sup> Session, 7-8 June 1983, 489 (Y4.G74/7:P93/1/1974).

[14] U.S. Congress. House. Committee on House Administration. *Legislative History of the Privacy Act of 1974*, S.3418 (*Public Law 93-579*): *Source Book on Privacy*. 94<sup>th</sup> Congress, 2<sup>nd</sup> Session, 1976, Joint Committee Print (Y4.G74/6:L52/3).

[15] Priscilla Regan, *Legislating Privacy: Technology, Social Values and Public Policy*. (Chapel Hill: University of North Carolina Press, 1995), 199.

[16] Aldrich, "Privacy Protection," 505-507.

[17] Commission of the European Community. *Communications on the Protection of Individuals in Relation to the Processing of Personal Data in the Community and Information Security*, Com (90)314.SYN 287, 44.

[18] Sarah Ellis and Charles Oppenheim. "Legal Issues for Information Professionals, Part III: Data protection and the Media – Background to the Data Protection Act 1984 and the EC Draft Directive on Data Protection," *Journal of Information Science* 19 (1993):85.

[19] "Directive 95/46/EC of the European Parliament and of the Council of 24 October on the Protection of Individuals with Regard to the Processing of Personal Data and the Free Movement of Such Data," *Official Journal of the European Community* 23 November 1995, no.L281, 31.

[20] Rebecca Vesely. "Cop-friendly Approach to Handling Medical Data," *Wired News* 12 (September 1997) (URL <http://www.wired.com/news/news/politics/story/6824.html>)

[21] Peter B. Swire and Robert E. Litan, Avoiding a Showdown over EU Privacy Laws, Brookings Policy Brief, no. 29 (February 1998) (URL <http://www.brook.edu/comm/policybriefs/pb029/pb29.htm>)

[22] *Ibid*.

\* Paper presented at the IASSIST Conference, May 21, 1998, at Yale University, New Haven, Connecticut. Jean Slemmons Stratford, University of California, Davis, and Juri Stratford, University of California, Davis.







INTERNATIONAL ASSOCIATION FOR  
SOCIAL SCIENCE INFORMATION  
SERVICE AND TECHNOLOGY

• • • • •  
ASSOCIATION INTERNATIONALE POUR  
LES SERVICES ET TECHNIQUES  
D'INFORMATION EN SCIENCES  
SOCIALES

The International Association for Social Science Information Services and Technology (IASSIST) is an international association of individuals who are engaged in the acquisition, processing, maintenance, and distribution of machine readable text and/or numeric social science data. The membership includes information system specialists, data base librarians or administrators, archivists, researchers, programmers, and managers. Their range of interests encompasses hard copy as well as machine readable data

Paid-up members enjoy voting rights and receive the **IASSIST QUARTERLY**. They also benefit from reduced fees for attendance at regional

and international conferences sponsored by IASSIST.

**Membership fees are:**

Regular Membership. \$40.00  
per calendar year.  
Student Membership: \$20.00  
per calendar year.

Institutional subscriptions to the quarterly are available, but do not confer voting rights or other membership benefits.

Institutional Subscription:  
\$70.00 per calendar year  
(includes one volume of the  
Quarterly)

## Membership form

I would like to become a member of IASSIST. Please see my choice below:

Options for payment in Canadian Dollars and by Major Credit Card are available. See the following web site for details:

<http://datalib.library.ualberta.ca/iassist/mbrship2.html>

- ☐ \$40 (US) Regular Member
- ☐ \$20 Student Member
- ☐ \$70 Subscription (payment must be made in US\$)
- ☐ List me in the membership directory
- ☐ Add me to the IASSIST listserv

Please make checks payable,  
in US funds, to IASSIST and  
Mail to:

IASSIST,  
Assistant Treasurer  
JoAnn Dionne  
50360 Warren Road  
Canton, MI 48187  
USA

Name: \_\_\_\_\_

Job Title: \_\_\_\_\_

Organization: \_\_\_\_\_

Address: \_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_

City: \_\_\_\_\_

State/Province: \_\_\_\_\_

Postal Code: \_\_\_\_\_

Country: \_\_\_\_\_

Phone: \_\_\_\_\_

FAX: \_\_\_\_\_

E-mail: \_\_\_\_\_

URL: \_\_\_\_\_



Return Undelivered Mail To:

**ASSIST QUARTERLY**  
c/o Wendy Treadwell  
1758 Pascal St. North  
Falcon Heights, MN 55113  
USA



Serials Department(SERLIBS82186344)  
Univ of North Carolina-Chapel Hill  
CB #3938 Davis Library  
Chapel Hill NC 27514-8890  
U.S.A.